On tests and significance in econometrics

Keuzenkamp, H.A.; Magnus, J.R.

Published in:
Journal of Econometrics

Publication date:
1995

JOURNAL OF ECONOMETRICS

Editorial Policy:
The Journal of Econometrics is designed to serve as an outlet for important new research in both theoretical and applied econometrics. Papers dealing with estimation and other methodological aspects of the application of statistical inference to economic data as well as papers dealing with the application of econometric techniques to substantive areas of economics fall within the scope of the Journal. Econometric research in the traditional divisions of the discipline or in the newly developing areas of social experimentation are decidedly within the range of the Journal's interests.
The Annals of Econometrics form an integral part of the Journal of Econometrics. Each issue of the Annals includes a collection of refereed papers on an important topic in econometrics.

Submission Fee:
Unsolicited manuscripts must be accompanied by a submission fee of US$50 for authors who currently do not subscribe to the Journal of Econometrics; subscribers are exempt. Personal cheques or money orders accompanying the manuscripts should be made payable to the Journal of Econometrics.

Publication Information:
JOURNAL OF ECONOMETRICS (ISSN 0304-4076). For 1995 volumes 65 - 69 are scheduled for publication. Subscription prices are availabe upon request from the publisher. Subscriptions are accepted on a prepaid basis only and are entered on a calendar year basis. Issues are sent by surface mail except to the following countries where air delivery via SAL is ensured: Argentina, Australia, Brazil, Canada, Hong Kong, India, Israel, Japan, Malaysia, Mexico, New Zealand, Pakistan, PR China, Singapore, South Africa, South Korea, Taiwan, Thailand, USA. For all other countries airmail rates are available upon request. Claims for missing issues must be made within six months of our publication (mailing) date. Please address all your requests regarding orders and subscription queries to: Elsevier Science S.A., P.O.B. 564, CH-1001 Lausanne 1, Switzerland.
Fax: 41-21-3235444.

ELSEVIER

Journal of Econometrics 67 (1995) 5–24

JOURNAL OF Econometrics

On tests and significance in econometrics

Hugo A. Keuzenkamp*[,a], Jan R. Magnus[b,c]

[a] *Department of Economics, Tilburg University, 5000 LE Tilburg, The Netherlands*
[b] *London School of Economics, London, UK*
[c] *CentER for Economic Research, Tilburg University, 5000 LE Tilburg, The Netherlands*

Abstract

Different aims of testing are investigated: theory testing, validity testing, simplification testing, and decision making. Different testing methodologies may serve these aims. In particular, the approaches of Fisher and Neyman–Pearson are considered. We discuss the meaning of statistical significance. Significance tests in the *Journal of Econometrics* are evaluated. The paper concludes with a challenge to ascertain the impact of statistical testing on economic thought.

Key words: Statistical tests; Inference; Significance level
JEL classification: B23; B40; C12

1. Introduction

In a provocative paper, McCloskey (1985, p. 182) contends that 'no proposition about economic behaviour has yet been overturned by econometrics'. McCloskey is not a lonely skeptic. Many outsiders are doubtful of the value added of econometric testing (e.g., Hahn, 1992). But also many econometricians are increasingly worried about the credibility gap between econometric theory and applied economics (for example, Spanos, 1986, p. 660). Whether the skeptics are right or wrong, we must face the question: What is the significance of testing econometric hypotheses?

* Corresponding author.

We are grateful to Michael McAleer and Mark Steel for their helpful suggestions.

SSDI 0304-4076(94)01624-9

Testing hypotheses belongs to the basic pastimes of econometricians. It is a compulsory topic in any course in introductory statistics and econometrics. In such a course, students are made familiar with notions like Type I and Type II errors, significance level, and power. This is firmly in the tradition of statistical testing along the lines proposed by Jerzy Neyman and Egon Pearson (1928, 1933). However, econometric practice seems closer to the approach of Sir R.A. Fisher, although he is rarely mentioned (apart from references to the F-test). We will clarify the differences between both approaches below.

At first sight, the lessons of an introductory econometrics course seem most useful, if one judges the amount of papers in economic journals that contain statistical tests. A casual investigation of titles of papers shows that there is a lot of 'testing' in the literature. Less comforting is the amount of 'evidence' that is found. What proportion of the results of tests, or of the evidence, is regarded to be powerful by a significant part of the audience? If the value added of testing is low, some reflections on the merits of testing in econometrics are due. It must be admitted that it is hard (but perhaps not impossible) to find a convincing example of a meaningful economic proposition, that has been rejected (or definitively supported) by econometric tests. Many statistical hypotheses have been tested and rejected. But in how many cases did the result remain unchallenged by a respectable colleague, or how often was a statistical rejection rather than common sense responsible for exorcizing a defective economic argument? If the value added of testing is low, some reflections on the merits of testing in econometrics are due.

In Section 2, we discuss aims of testing, relating them to popular views in the philosophy of science. In Section 3, some statistical methods of testing are discussed. Statistical significance is analyzed in Section 4, while testing in the *Journal of Econometrics* is the topic of Section 5. We conclude the paper with a challenge to the readers.

2. Aims of testing

Why test? Sometimes one wonders about the abundance of tests reported in empirical papers, as the purpose of many of these tests is not always communicated to the reader. Occasionally, the number of test statistics reported in a paper exceeds the number of observations used in calculating them! In many cases, the implications of a positive or negative result are not made clear. If a null hypothesis that apes behave perfectly rationally is rejected at the 5% significance level, do we care? And should we be interested in the normality of the residuals, or would it be more useful to put the tests aside and read Darwin's *Origin of Species* instead? But perhaps it is inherent to our occupation as econometricians that we stick to providing statistical inferences.

An important reason for the popularity of testing is that it is often thought to be a major if not the main ingredient to scientific progress (Popper, 1968; Stigler, 1965, p. 12; Blaug, 1980) and the best way to move from alchemy to science (remember Hendry's three golden rules of econometrics: test, test, and test; *cf.* Hendry, 1980). According to Popper's demarcation criterion, scientific hypotheses are falsifiable ones. Unfalsifiable propositions belong to the domain of metaphysics, not science. You want to be scientific? Then test your hypotheses! And one of the founders of statistical testing writes: 'Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences' (Fisher, 1973a, p. 2).[1] Hence, if we wish to be scientific, then let us test hypotheses — statistical hypotheses, that is.

Occasionally, econometricians reach out to the scientific ideal of testing economic hypotheses, confronting theory (more precisely, a particular specification of the theory) with facts. We will call this *theory testing*. It is the most ambitious of the aims of testing. Examples are testing monetarism, real business cycle theory, the efficient market hypothesis, hysteresis, or rational consumer behaviour. Ideally, tests in this category deal with efforts to test one theory against a rival one, that is, to discriminate (monetarism versus Keynesianism, hysteresis versus heterogeneity). Scientific progress, it is often argued, consists of replacing a defective theory by a better one. Nonnested hypotheses tests, encompassing tests, but also specification tests and, occasionally, model selection tests, belong to the [illegible] of theory testing.

Theory testing is closely related to a once popular approach in the philosophy of science, the hypothetico-deductive (HD) method.[2] This method consists of formulating sets of hypotheses, from which predictions of novel facts can be deduced: the consequences. These are the testable implications. The empirical scientist either should try to measure the degree of confirmation (according to logical positivists of the Vienna Circle, like Carnap) or try to falsify these testable implications (according to falsificationists like Popper). Prominent members of the Cowles Commission, in particular Haavelmo (1944) and Koopmans (1947), advocated an HD approach to econometrics which resulted in a formalistic methodology of economic inference. More recently, HD econometrics can be found in the writings of new classical economists, in particular by those who search for 'deep' (structural) parameters. Another recent publication in the tradition of the HD approach is Stigum (1990). But Summers (1991) forcefully

[1] Ironically, the quote continuous as follows: 'This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth methods adequate to the treatment of economic data, in so far as they exist, have mostly been developed in the study of biology and the other sciences'.

[2] See Chapter 3 in Earman (1992) for discussion and references.

argues that formalistic empirical econometrics has not yielded interesting insights in macroeconomics: this approach to inference leads merely to a 'scientific illusion'.

Popper's falsificationism has had a strong impact on the minds of economists. Popper is about the only philosopher of science occasionally quoted in *Econometrica*. In the philosophy of science literature, however, falsificationism has become increasingly unpopular. Not the least because actual science rarely follows the Popperian maxims. As Hacking (1983, p. 15) notes, 'accepting and rejecting is a rather minor part of science' (see also the contributions in Earman, 1983, and those in De Marchi, 1988). Theory testing is an aim that, in practice, is less important than some would like to think.

An alternative to hypothetico-deductivism is Bayesian inductive inference. Carnap (1952) also contributed to this approach, but Jeffreys (1961) had a stronger impact on the Bayesian minority in econometrics.[3] This alternative approach shares with the HD method a belief in growth of knowledge (a feature that has been attacked by so-called post-modernist philosophy; see Mirowski, 1995). However, the aim of theory testing is less important in the Bayesian inductive tradition than within Popperian hypothetico-deductivism (see, e.g., Leamer, 1978, p. 9). Some Bayesians do not see merit in hypothesis testing, they hold measurement as the more interesting aim of inference. If rival hypotheses exist and, e.g., prediction is the purpose of inference, the best one can do is to weigh the alternative hypotheses and use a basket of weighed predictions. Other Bayesians use Jeffreys' Posterior Odds Ratios as a test statistic. If decision making is the purpose of the test, then the behaviouristic approach of Savage (1972) is advocated by some Bayesians (below, we will discuss decision making as a distinct aim of testing).

If theory testing is an interesting aim at all, it is not yet clear that econometrics is the best tool for this purpose. Identifying informative historical episodes (see, e.g., Summers, 1991) or devising laboratory experiments (increasingly popular among game theorists, who rarely supplement their experiments with statistical analysis, as casual reading of such experimental reports in *Econometrica* reveals) may generate more effective tests than many Uniformly Most Powerful (UMP) tests. Consider Science with capital S: physics. Here, sophisticated statistical considerations play a minor role in appraising theories. Giere (1988, p. 190) discusses the different attitudes towards data appraisal in nuclear physics and the social sciences. Nuclear physicists tend to judge the fit between empirical and theoretical models primarily on qualitative arguments. Test statistics such as χ^2 are rarely reported in nuclear physics papers contained in, e.g., the *Physical Review*.[4] Theory (or hypothesis) testing does not necessarily depend upon the tools we learned in our statistics courses.

[3] See Howson and Urbach (1989) and Earman (1992) for philosophical backgrounds of Bayesian confirmation theory.

[4] Baird (1988) makes a similar observation.

We will now turn to other aims of testing, less prominent in philosophical writings, but dominant in practical research. Most tests are not as ambitious as the theory tests discussed above. An important case is the class of the (*statistical*) *validity tests* (misspecification tests or diagnostic checks). Validity tests are performed in order to find out whether the statistical assumptions underlying some model are credible. Spanos (1995) is an example of extensive validity testing. He follows the argument that in order to pursue a theory test, one first has to be sure of the validity of the statistical assumptions that are made. According to this view, validity testing is a prerequisite to theory testing (note that Granger et al., 1995, advocate the reverse ordering). If theory testing is not the ultimate aim, validity testing still may be important. Much empirical work aims to show that a particular model (formally or informally related to some theory) is able to represent the data. If much information in the data remains unexploited (for example, revealed by non-white-noise residuals), this representation will be suspect or unconvincing to a large part of the audience.

Sometimes, however, it is argued that the merits of validity tests should not be over-emphasized. One may obtain a very neat 'valid' statistical model of some economic phenomenon, after extensive torturing of the data. Such a specification suggests much more precise information than the data actually contain. Sensitivity analysis, either along the lines of Leamer (1978) or Friedman (see, for example, the discussion in Summers, 1991), is at least as important as validity testing in order to make credible inferences. Illuminating in this context is the exchange between Hendry and Ericsson (1991) and Friedman and Schwartz (1991).

A third important aim of testing is *simplification testing*. Simple models that do not perform notably worse than more complex ones are typically preferred to the complex one. Inference conditional on exogeneity assumptions is often preferred to full information estimation. Still, it is regularly argued that, apart from convenience, there are no clear formal reasons why simple models deserve special credit (but see Keuzenkamp and McAleer, 1995, for discussion and further references). A popular view on simplification testing is that the researcher should start with a very general model, and perform a downward test strategy in which uninformative elements of a model are deleted (Hendry and Ericsson, 1991). In practice, many researchers feel that simplicity matters, but rather than testing from general to simple, they perform iterative simplification searches.

Finally, a frequently expressed goal of testing is *decision making* (e.g., Granger et al., 1995). This view on testing, and its implementation to statistics, is primarily due to the Neyman–Pearson theory of inductive behaviour (Neyman and Pearson, 1928, 1933). The decision-theoretic approach to testing has been further elaborated by Wald and, from a Bayesian perspective, by Savage (1972). Lehmann (1986) is the authorative reference for the frequentist approach, while Berger (1985) provides the Bayesian arguments.

Decision making, based on statistical acceptance rules, can be important for process quality control, but may even be extended to the appraisal of theories. This brings us back to theory testing. Lakatos, the neo-Popperian philosopher, claims that the Neyman–Pearson version of theory testing 'rests completely on methodological falsificationism' (Lakatos, 1978, p. 25n). Apart from the fact that this reverses historical priority (the first German edition of Popper, 1968, appeared in 1934), it is also at odds with Popper's own rejection of behaviourism (see Keuzenkamp, 1994, Ch. 3.4.4, for further discussion). Still, it may be argued that the Neyman–Pearson approach to theory testing (popularized in econometrics by Haavelmo, 1944) fits in the broader hypothetico-deductive approach, of which Popper's version is only one brand.[5]

At this point, one of the most bitter disputes in science deserves special mention: the Fisher versus Neyman–Pearson controversy. One of the sources of their dispute was the aim of testing. While Neyman–Pearson acceptance rules can be placed in the hypothetico-deductive camp, the views of Fisher are closer to a Bayesian inductive approach.[6] Fisher's theory of estimation and testing is a theory of learning, meant for inductive *inference* from small samples. Neyman and Pearson opposed aiming at inductive inference. They interpret tests along behaviouristic lines, as acceptance rules in the context of repeated sampling. At best, Fisher was willing to support such an interpretation for problems in commerce or technology, but not for appraising scientific hypotheses. The reason is that in such cases, repeated sampling is a misleading fiction, and there is no well-defined decision problem. Many advances made in science do not serve a well-specified purpose, moreover, 'they may be put sooner or later to the service of a number of purposes, of which we can know nothing' at present (Fisher, 1973b, pp. 106–107). Even if there would be a well-specified decision problem, estimation was of more interest to Fisher than devising UMP tests.[7] Indeed, for many econometric papers that appear in the *Journal of Econometrics* and *Econometrica* among others, it is hard to define the decision problem and loss functions that should figure in the background if a Neyman–Pearson approach were followed.

Such doubts are shared by Savage (1972, p. 254) who writes that, although having tested many sharp null hypotheses, he is unable to give a satisfactory analysis of testing such hypotheses. To him, the role of extreme null hypotheses

[5] Giere (1983) is a philosopher's view on theory testing, which is an augmented version of the Neyman–Pearson theory (without mentioning Neyman–Pearson).

[6] Although Fisher rejected Bayesianism in cases where there is no informative prior probability, he had an alternative, so-called fiducial inference (see Fisher, 1973b). This has been characterized as 'a bold attempt to make the Bayesian omelette without breaking the Bayesian eggs' (Savage, 1961, p. 578).

[7] For Fisher's views on the Neyman–Pearson methodology, see Fisher (1973b, pp. 42, 80, 103–107) and Section 3 below.

in science is 'obscure'. A problem with such hypotheses is that in many cases the loss associated with the alternative is zero, only a loss (or gain) exists if the null is exactly satisfied. The behaviouristic theory of inference is difficult to apply in such circumstances. Still, many econometricians do test sharp null hypotheses, and think that these tests are straightforward applications of testing in the Neyman–Pearson tradition.

Many such sharp null hypotheses are of little scientific interest anyway. Still, even the best journals, such as the *Journal of Econometrics*, report tests of purchasing power parity or perfectly efficient markets, even if we are all aware that these theories are not literary true. Would it not be more interesting, in such cases, to measure how close the real world is to the ideal world of the theories? According to Leamer (1978, p. 9), hypothesis testing searches are rare, while Jeffreys (1961, p. 389) remarks that 'what are called significance tests in agricultural experiments seem to me to be very largely problems of pure estimation'. Jeffreys' argument, if applied to economics, would run like this. A labour economist has a very good idea of what to expect when estimating a model that analyzes the returns to schooling. His problem is to choose the variables, and obtain a sample of sufficient size, such that the effect of education and other variables of interest become detectable. It is the magnitude of the effects that is of primary interest. Any level of significance can be obtained by making the sample size large enough, unless the null hypothesis is exactly true (Berkson, 1938).

This concludes our discussion of four distinct aims of testing: theory testing, validity testing, simplification testing, and testing for making decisions. We now turn to a number of statistical methods that serve these aims of testing.

3. Methods of statistical testing

Informal statistical testing of hypotheses has a long history (frequently cited examples of significance testing *avant la lettre* are Arbuthnot on male vs. female births in 1710, Mitchell on the distribution of stars in 1767, and Laplace in 1812; see, e.g., Hacking, 1975, p. 168; Baird, 1988). In 1885, Edgeworth introduced the term 'significant' in statistics (Baird, 1988). The modern approach to significance testing starts with Karl Pearson's goodness of fit test for large samples (Pearson, 1900). The basic philosophy of his testing procedures is as follows. A sample is used to estimate a test parameter of interest. The distribution under the null is known, and if the estimate falls too far into the tail of the distribution, one of the following conclusions must be drawn: either something very uncommon has happended or the null hypothesis is wrong. The P-value (tail area integral) thus obtained is compared with a benchmark, like 0.01 or 0.05 (see Section 4 below). Subsequently, in statistical inference, the option that the null is wrong is chosen if the P-value falls beyond the benchmark level.

Henry L. Moore belonged to the first economists who applied Pearson's methods to economics (see also Stigler, 1965). W.S. Gosset, better known by his pseudonym Student, introduced the small-sample t-test for the equality of means in 1908. But Fisher greatly extended the scope of testing, he also derived the correct degrees of freedom that belong to different applications of the tests.[8] Fisher also invented analysis of variance and the F-test (originally labelled as z-test).[9] His method of maximum likelihood remains widely used, but his reliance on the likelihood principle and conditional inference is not generally accepted.[10] In his doctoral thesis, Koopmans (1937) built on Fisher's methods of estimation and testing. It is notable that Tinbergen (1939), who supervised Koopmans' thesis, does not make use of significance testing. Instead, Tinbergen's approach is better characterized as *importance testing*. (Jeffreys would probably argue that this again is a problem of estimation, rather than testing.)

To sum up the Fisherian theory of significance testing, it contains the following characteristics:

(i) reliance on tail areas (P-values),
(ii) intended for small samples,
(iii) instruments for inductive scientific inference.

This approach has received two kinds of criticism. The first comes from Jeffreys, who rejects (i). The second criticism is given by Neyman and Egon Pearson (the son of Karl Pearson), who argue that (i) alone is not sufficient to select a test procedure, (ii) should be replaced by repeated sampling, and who also disagree with (iii). To start with Jeffreys (1961, p. 385), he argues that any particular set of observations has a low probability to obtain. Hence: 'If mere improbability of the observations, given the hypothesis, was the criterion, any hypothesis whatever would be rejected'. In the posterior odds approach, advocated by Jeffreys, this problem vanishes since the ratio of probability values for two distinct hypotheses will be informative; the small factors cancel. The P-integral methodology instead does not appraise the probability of the actual observations, in view of a hypothesis, but takes the observations that would generate P-values beyond the benchmark level. 'The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure.' (Jeffreys, 1961, p. 385, his italics). In other words, this method of inference violates the likelihood principle. Another criticism of Jeffreys (1961, p. 390) is that it is not very useful to reject a hypothesis without having some idea of what to put in its place (see also Keuzenkamp and Barten, 1995).

We already noted that the test approach advanced in Neyman and Pearson (1928) and further explored in their later writings (in particular, Neyman and Pearson, 1933) diverges from Fisher's in some important respects (but note that Neyman and Pearson adopted many of Fisher's insights, and at first were even convinced that their work was just an improvement of Fisher's; see Fisher Box, 1978, for details, and Reid, 1982, for a perspective that sides with Neyman and Pearson).

First, there is a philosophical distinction between Neyman–Pearson methods and Fisherian hypothesis testing. The Neyman–Pearson approach is not based on inductive aspirations (such as Fisher's), but is directed to behaviour, following the then fashionable behaviouristic school of thought in psychology and other disciplines (J.B. Watson's classic on behaviourism appeared in 1930, and papers on that topic have appeared since 1913). The Neyman–Pearson tests are acceptance procedures, decision rules (see above), not methods of inference.

Secondly, Neyman and Pearson were dissatisfied with the existence of a wide range of tests while no one knew which one was the 'best'. According to Fisher, the research worker normally knows what alternatives are relevant (without specifying them) and, therefore, what test is to be selected. However, Neyman and Pearson tried to define general optimality conditions for tests, in a context of repeated sampling. This can only be done after the unspecified alternative hypotheses in the Fisher approach are replaced by specific alternatives. Once this step is made, the notion of errors of the first kind (rejecting a correct null hypothesis) and the second kind (accepting a false null hypothesis) can be introduced, the power of a test is defined, and UMP tests can be obtained in a number of cases. The likelihood ratio (LR) test (first proposed on intuitive grounds in 1928, then justified on theoretical grounds in 1933), stands out as their principal contribution to the theory of hypotheses testing.

Summarizing, to contrast the Neyman–Pearson approach to Fisher's, the following points characterize the Neyman–Pearson methodology:

(i) emphasis on size and power, leading to UMP tests,
(ii) applications to contexts of repeated sampling,
(iii) instruments for inductive behaviour and decision making.

At the formal level Neyman and Pearson seem to have won the battle with Fisher. Many economists have learnt Neyman–Pearson methods of hypothesis testing in their first introductory course in statistics. For example, the treatment of hypothesis testing in the popular statistics textbook of Wonnacott and

[8] In 1922, Gosset sent his tables of the t-test to Fisher, writing 'you are the only man that's ever likely to use them!' See Joan Fisher Box (1978, pp. 116, 454).

[9] z is a transformation of F which was easier to interpolate ($z = \frac{1}{2}\ln F$).

[10] Lehmann (1993) discusses the issue of conditional inference and compares Fisher's perspective with that of Neyman.

Wonnacott (1985) is based on a simplified version of Neyman–Pearson testing (an explicit reference to Neyman–Pearson is given on p. 257).[11] Another popular textbook used in econometrics is Judge et al. (1988). Hypothesis testing is interpreted as a decision problem in the light of the costs of making an incorrect decision (p. 93), the discussion is entirely in the spirit of Neyman–Pearson procedures. The encyclopaedic nature of this book is reflected in alternative discussions of hypothesis testing, in particular posterior odds (p. 131), but Fisher's approach is not discussed. Goldberger's (1991) textbook deals with hypothesis testing from a Neyman–Pearson perspective (Chapter 20) and even explains the Neyman–Pearson implication of a rejection of a null at a 5% significance level: 'Loosely speaking, when the null is true, in 5% of the samples drawn from the population, the decision will be "reject the null"' (Goldberger, 1991, p. 215). Finally, the survey paper of Engle (1984) gives an overview of test procedures (Lagrange Multiplier tests, starting at the null and testing whether movements to the alternative lead to an improvement, Wald tests, starting at the alternative, and LR tests that may proceed symmetrically), all based on Neyman–Pearson principles.

The implementation of Neyman–Pearson methods at the practical level is not easy, though. There is a wide divergence between empirical econometrics and the maxims of a 'celibate priesthood of statistical theorists', as Leamer (1978, p. vi) observes. One reason for the dominance of the Neyman–Pearson approach among this priesthood might be that it lends itself to mathematical recreation. Another nonsubstantive reason is the attraction that the words 'best' and 'powerful' exert. But it is more interesting to evaluate the substantive features of Neyman–Pearson testing. They have several drawbacks.

First, consider the notion of power. According to Fisher, emphasis on power is in many cases hardly relevant. To a practical researcher, 'it is, of course, a matter of indifference with what probability he might be led to accept the hypothesis falsely, for in his case he is not accepting it' (Fisher, 1973b, p. 42). Another problem with the power of a test is that it may be low when the model is misspecified (i.e., the maintained hypothesis is wrong). On the other hand, some tests (such as the Durbin–Watson test) happen to be rather powerful against misspecifications for which they are not intended. The Neyman–Pearson approach hinges on the 'axiom of correct specification' (Leamer, 1978, p. 4). Recently, efforts have been made to extend the scope of Neyman–Pearson methods to misspecified models (an example is Vuong, 1989). An alternative, becoming increasingly popular, is to use nonparametric methods of inference. Some investigators who support this approach believe that they can avoid

[11] In a subtle deviation from the Neyman–Pearson interpretation of testing, Wonnacott and Wonnacott (1985, p. 259) argue that a statistical test is a device to judge the acceptability or plausibility of the hypothesis.

making the specification errors that afflict parametric inference (see also Härdle and Kirman, 1995).

Secondly, the fiction of repeated sampling is questionable. One of the first critics was Fisher. He states that 'if we possess a unique sample in Student's sense on which significance tests are to be performed, there is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard our sample as belonging: so that the phrase "repeated sampling from the same population" does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination' (Fisher, 1955, p. 71).[12]

Thirdly, although some argue that the decision-theoretical approach should be natural to economists, in many cases it is very difficult to determine what decision really inspires a particular test and what loss is involved (see Section 2 above). Although the decision-theoretical approach to theory testing is obscure, it may be helpful in cases of validity testing, which has some resemblance to process quality control (if we are willing to ignore the Neyman–Pearson emphasis on repeated sampling). The loss, e.g., involved with serial correlation, might be that readers who stick to the 5% convention will stop reading a research report if they suspect that serial correlation is not properly taken care of. It still is not a formal loss, expressed in dollars, but loss resulting from loss of readership driven by (bad or good) conventions (see the blunt comments in Friedman, 1988, footnote 11). Although such a justification of Neyman–Pearson methods for validity testing could be sustained, one of the proponents of Neyman–Pearson methods makes a distinction between 'model design criteria (exhausting the available data evidence) and genuine tests in the Neyman–Pearson sense (based on previously unavailable evidence)' (Hendry, 1992, p. 366). He adds that 'only information that arrived after a model is in the public domain can be deemed an adequate basis for a test' (Hendry 1992, p. 374). If we understand Hendry correctly, he argues that validity testing (model design) is not genuinely Neyman–Pearson, but theory testing is a kind of Neyman–Pearson quality control test. We agree with the first statement, as genuine Neyman–Pearson testing requires repeated sampling. For the same reason, the second statement seems less convincing (the accumulation of a handful of extra quarterly observations can hardly count as an instance of repeated sampling). Moreover, it is not clear how to interpret scientific inference as a genuine decision problem, to be solved with behaviouristic arguments.

A fourth problem with Neyman–Pearson testing is that, if we have two explicitly specified alternatives to choose from, it is more natural to choose the one with the higher likelihood without considering the power functions and

[12] The reference to Venn relates to the second edition of *The Logic of Chance*.

without having to take one as the null and the other as the alternative (see Jeffreys, 1961, p. 396; a Bayesian would consider the posterior odds ratio). Vuong (1989) discusses how the LR test can be used in a symmetric way for model selection and testing nonnested hypotheses in a context of independent observations.

A general problem of significance testing, whether Neyman–Pearson or Fisherian, occurs when multiple tests are carried out. Depending on how dependent these tests are, the overall significance level may be much higher than the individual significance levels. The problem was recognized by the early econometricians. Indeed, Haavelmo (1944, p. 83) already discusses, in today's parlance, pre-testing. It is valid, he argues, but not if the set of a priori admissible hypotheses is a 'function of the sample point'. This rules out experiments with the maintained hypothesis. Naive induction, as one might call this method, cannot be totally ignored (to use an understatement) if one appraises empirical econometrics. Moreover, not only the maintained hypothesis may be the result of 'data mining', but not infrequently the alternative hypothesis is inspired by a rejection of a null rather than specified in advance, as it should in case of the Neyman–Pearson methods. The problem of interpreting the resulting test statistics remains unsolved today (see Godfrey, 1988, p. 3; Leamer, 1978, p. 5). Indeed, as Hendry (1992, p. 369) notes, test statistics can frequently be made insignificant by construction, since the residuals are not autonomous but derived processes.

Many authors agree that significance tests are not the only or ultimate tests of economic hypotheses. Friedman is not alone in his verdict that 'the real test of a theory' lies in its predictive ability, a theme he has consistently repeated since his 1940 review of Tinbergen's (1939) statistical (importance) tests of business cycle theories. This ability may be evaluated quantitatively, with statistical tools, but also qualitatively. Theil (1971, p. 545) argues that statistical procedures are not sacrosanct in modelling: 'The real test is provided by prediction based on an independent set of data. It is not at all self-evident that selections that are exclusively based on the smallest residual-variance estimates lead to the best predictions'. Similar opinions are expressed by Hendry (1992, p. 374), Zellner (1988, p. 31), and numerous other econometricians.

4. What is 'significant'?

If economists have natural constants, then the most well-known is 0.05. From early applications to the most recent hypothesis tests, investigators have relied on a significance level of 0.01 or 0.05. This convention owes much to Fisher's tabulation of statistical distributions in Fisher (1973a), first published in 1925.[13] Fisher and Gosset ('Student') cooperated in calculating tables for the

[13] We are grateful to Jim Durbin for historical advice on this matter. See also Hall and Selinger (1986) for a discussion of the historical roots of the 5% convention.

t-distribution. Fisher also tabulated the distributions of χ^2 and the z-transformation of the F-distribution. Originally, Fisher hoped to include existing tables of χ^2, made by W.P. Elderton and published in *Biometrika* of 1902, in his book. However, Karl Pearson (editor of *Biometrika*, father of Egon Pearson) did not allow him to reprint those tables. Pearson did not approve of Fisher's refinements of interpreting the χ^2 test (in particular, the issue of degrees of freedom), and their personal relations were bad (see M.G. Kendall, 1963; Fisher Box, 1978). Hence, Fisher was forced to make a distinct table by himself. He decided to turn the tables inside out, which seemed more convenient as well. Existing tables provided P-values (tail areas) for given values of χ^2 and t. Fisher argues: 'Instead of giving the values of P corresponding to an arbitrary series of values of χ^2, we have given the values of χ^2 corresponding to specially selected values of P' (Fisher, 1973a, p. 79; see Fisher Box, 1978, pp. 246–247, for further background). The P-values for which the χ^2 distribution was tabulated (for $n = 1, \ldots, 30$) are 0.99, 0.98, 0.90, 0.80, 0.70, 0.50, 0.30, 0.20, 0.10, 0.05, 0.02, and 0.01 (Fisher, 1973a, pp. 112–113). A similarly extensive tabulation is provided for the t-distribution (*op. cit.*, p. 176). Hence, the 0.05 significance level is not singled out as one with special merit, although Fisher (1973a, pp. 114–115) writes: 'If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance; this is roughly equivalent to the corresponding limit $P = 0.05$ already used for the χ^2 distribution.' Finally, as the z (or F) distribution needs separate tables for all significance levels, Fisher decided to tabulate this distribution for 'three especially important values of P' (Fisher 1973a, p. 228, pp. 244–249): 0.05, 0.01, and 0.001. Those are the significance levels that we observe as the few natural constants that economists rely on when they do empirical research. Still, Fisher (1973b, p. 42) warns against dogmatically applying a fixed level of significance in all circumstances.

Although Fisher was not the first statistician who tested at a 5% significance level, he facilitated its breakthrough by suggesting to use 'significant' as an abbreviation of 'significant at the 5% level' and moreover by means of his convenient tabulation. His interest in small sample analysis is reflected by the fact that his tables run from $n = 1$ to $n = 30$ (and in some cases also include 60 and infinity). Fisher does not discuss what the appropriate significance levels are for large samples. Berkson (1938) observed that, as sample size grows to infinity, any sharp null hypothesis is likely to be rejected at a fixed significance level. This has yielded the suggestion to vary the significance level with sample size (Leamer, 1978). As we will see in the next section, this suggestion has been largely ignored in practice. A possible explanation is that statisticians try to measure parameters using some benchmark level of precision. A fixed significance level serves this purpose. If our conjecture is valid, we expect to find that models estimated with many observations to have a higher-dimensional parameter vector (we did not attempt to test this hypothesis statistically).

Given the conventional significance levels, it remains to explain what they really mean. Most textbooks ignore this issue, but there are notable exceptions. Wonnacott and Wonnacott (1985) prefer the expression 'statistically discernible at the 5% error level' to the more familiar phraseology 'statistically significant at the 5% significance level'. An explicit warning is given that statistical significance is not the same as importance (or substance, in Goldberger, 1991, p. 240). Furthermore, the discussion of the χ^2 test points to some limitations of hypotheses tests (Wonnacott and Wonnacott, 1985, pp. 488–489), in particular to the fact that such tests often give answers to the wrong question. Goldberger's (1991, p. 215) explanation of the meaning of rejection at a 5% significance level, quoted in Section 2 above, is the valid interpretation of Neyman–Pearson testing. But how often is the question that an econometrician has to answer a decision problem in the context of repeated sampling? Fisher's interpretation of a small *P*-value (which follows the tradition of Laplace to K. Pearson), that either something very unlikely has happened or the null is false, may be more useful in econometric practice. A third alternative is to interpret *P*-values as odds factors. In this case, however, the Bayesian (posterior odds) perspective may be preferred, as Jeffreys already showed that posterior odds ratios often tell a different story than significance tests based on a fixed significance level (Jeffreys, 1961; Berger, 1985).

A different perspective on interpreting significance tests arises when one realizes that, at least in economics, most inferences are based on extensive data mining. Karl Pearson objected to using arbitrary levels of significance to assess the validity of a hypothesis. Instead, statistics involves curve fitting and gradual approximation from poor fit to good fit, not from falsity to truth. Goodness of fit tests 'are used to ascertain whether a reasonable *graduation* curve has been achieved, and not to assert whether one or another hypothesis is true or false' (K. Pearson, letter to *Nature*, 1935, cited in Hall and Selinger, 1986, p. 359). This skeptical view on significance testing is not much heard today, Milton Friedman being one of the exceptions (see, e.g. Friedman, 1988, p. 323, footnote 11).

5. Testing in the *Journal of Econometrics*

The strong emphasis in journals on significance testing not only exists in economics. The *JRSS* (*Journal of the Royal Statistical Society*) is sometimes referred to as the *JSSR* (*Journal of Statistically Significant Results*).[14] Similarly, the economic literature abounds with significance tests. Zellner (1979) contains a small survey of 22 quantitative articles in five issues of different leading economic journals in 1978. He finds that significance testing is very popular, that

[14] See Wonnacott and Wonnacott (1986, p. 573), who also discuss the 'editor's bias' of preferring significant test results.

1% and 5% significance levels dominate, and power considerations are rarely discussed, despite the dominance of Neyman–Pearson methods in the training of economists. According to another survey, of Canterbery and Burkhardt (1983, p. 31), out of 542 empirical papers that appeared in the *American Economic Review*, *Journal of Political Economy*, *Economic Journal*, and *Quarterly Journal of Economics* from 1873–1978, only three articles attempted to refute the hypothesis under investigation. Although this may sound unnerving, there is a reason why econometricians do not play the falsificationist game with much enthusiasm. In most cases, rejection of economic hypotheses is easy, whereas verification is hard (anyone with experience in economic modelling knows how difficult it can be to obtain models that are 'satisfactory').

For the purpose of investigating the significance of significance tests, we surveyed the papers in the *Journal of Econometrics* (excluding the *Annals*), Volumes 1–46 (1973–1990). In total, 668 papers were counted. Of those papers, 17% have 'test' (or 'testing') in the title. Not all 668 papers contain data. 26% contain artificial data, used for Monte Carlo investigations. Of the papers containing empirical data, many use those data for the purpose of illustration only (this is obviously the case if, for example, 'Klein I' is estimated – one of the most popular models in this Journal). We excluded those papers from our analysis (in a few cases, the choice is somewhat arbitrary).

This left 137 papers (21%) with an empirical message that exceeds mere illustration. Among those papers, 99 made use of significance tests. The significance levels were (in increasing popularity):

0.02	(1 paper),
0.001	(2 papers),
0.10	(2 papers),
0.005	(5 papers),
0.01	(26 papers),
0.05	(63 papers).[15]

The choice of the significance level might depend on sample size, in view of Berkson's (1938) observation and similar recommendations of Jeffreys (1961, p. 435) and Leamer (1978, p. 105). Hence, we investigated the relation between significance level and sample size.[16] Indeed, a few explicit references can be found concerning sample size and the trade-off between 'expected loss from

[15] A number of papers refers to more than one significance level. In that case, the most stringent (lowest) level is reported. Where this could be verified, it turned out that papers where the significance level remains implicit ('the parameter is significant') all refer to the 5% significance level. Hence, in those cases where this could not be verified, we assume that the 5% level is applied as well. Papers which report *P*-values are not included in this count.

[16] As it is not our purpose to blame specific authors, we refer in the following only to volume numbers and not to specific papers.

Type I and Type II errors', as in a paper in Vol. 22 where a sample of 728 observations inspires a 1% significance level. In a paper in Vol. 44, a null hypothesis is rejected at a 5% significance level which the author had rather preferred to accept. Given sample size (14,487 observations), the author argues that conventional significance levels are not appropriate. Instead, with such large samples, 'a case can be made for using a Bayesian procedure'. However, upon further analysis of the relation between significance level and sample size in empirical papers, it appears that the correlation between sample size and significance level is opposite to what might be expected. The correlation coefficient is positive and has a value of 0.2! Hence, in practice, the choice of significance levels seems arbitrary and depends more on convention and, occasionally, on the desire of an investigator to reject or accept a hypothesis rather than on a well-defined evaluation of conceivable losses that might result from incorrect decisions.

The papers which explicitly attempt to test a theory statistically are rare (less than a dozen); the cases where a clear conclusion (acceptance or rejection of the theory) emerges, are even rarer. In cases where a decisive conclusion is obtained, the same volume may contain a test of the same hypothesis with the opposite result (e.g., tests of efficient markets in Vol. 4). If a theory is rejected (e.g., neoclassical production theory, Vol. 7, or the theory of demand, Vol. 15, both at a 1% significance level), it often remains unclear what the implications are (the 'not very constructive conclusion' is 'worth remembering' or 'rejection of the theory is not necessarily implied'). Occasionally, it is acknowledged that the implication of significance tests is often unclear: a rejection does not necessarily mean a rejection of the hypothesis of interest, as auxiliary hypotheses might be false instead (see also Keuzenkamp and Barten, 1995).

In analyzing significance tests published in the *Journal of Econometrics*, we were surprised that some elementary rules are occasionally violated. Sample sizes are not always reported. Investigators claim to test a hypothesis at an unspecified or a 95% significance level, when a 5% level is meant. The advice by Goldberger (1991, p. 217), to use correct wording, is appropriate not only to undergraduate students.

6. Theory testing and significance: A challenge

According to Engle (1984, p. 776): 'If the confrontation of economic theories with observable phenomena is the objective of empirical research, then hypothesis testing is the primary tool of analysis.' This is the view of a mainstream econometric theorist. This view puts high emphasis on testing, but many econometricians are aware of the limited impact of testing and concur in McCloskey's skepticism. Spanos (1986, p. 660) acknowledges that 'to my knowledge, no economic theory was ever abandoned because it was rejected by some empirical econometric test, nor was a clear-cut decision between competing theories made in lieu of the evidence of such a test'. The same verdict has been expressed by economic theorists like Hahn (1992): 'I know of no economic theory which all reasonable people would agree to have been falsified.' Not only theorists argue like this. Summers (1991, p. 133) writes: 'It is difficult to think today of many empirical studies from more than a decade ago whose bottom line was a parameter estimate or the acceptance or rejection of a hypothesis.' In many cases, formal econometric hypothesis testing is unpersuasive. The value added of econometric tests may be less than desired. Some even argue that we only test what we already believe beforehand. According to Keynes (1921): 'The truth is that sensible investigators only employ the correlation coefficient to test or confirm conclusions at which they have arrived on other grounds. But that does not validate the crude way in which the argument is sometimes presented, or prevents it from misleading the unwary – since not all investigators are sensible.'

These skeptical observations on significance testing for the purpose of theory testing should challenge econometricians who think otherwise. Therefore, we invite readers to name a paper that contains significance tests which significantly changed the way economists think about some economic proposition. The following rules of the game apply:

1. You may interpret the notion 'significance test' broadly, i.e., both Fisher's and Neyman–Pearson's interpretations are accepted (please indicate if one of those interpretations is most appropriate).
2. Give exact reference to author, paper, journal, etc., and to the particular test(s) you think persuaded economists.
3. Summarize the test result by (a) what the hypothesis tested is, (b) whether the hypothesis is accepted or rejected, and (c) if the hypothesis is rejected, is there a constructive message?
4. If possible, provide auxiliary evidence that the particular test has been persuasive to others.

The responses to this challenge will be processed statistically and if the results are of sufficient interest, they will be reported. You may send us your suggestion until six months after publication of this Issue. The most convincing contribution will be awarded with an invitation for a one week visit to CentER (expenses paid).

References

Baird, Davis, 1988, Significance tests, history and logic, in: Samuel Kotz and Norman L. Johnson, eds., Encyclopedia of statistical sciences, Vol. 8 (Wiley, New York, NY).

Berger, James O., 1985, Statistical decision theory and Bayesian analysis, 2nd ed. (Springer Verlag, Berlin).
Berkson, J., 1938, Some difficulties of interpretation encountered in the application of the chi-squared test, Journal of the American Statistical Association 33, 526–542.
Blaug, Mark, 1980, The methodology of economics or how economists explain (Cambridge University Press, Cambridge).
Canterbery, E. Ray and Robert J. Burkhardt, 1983, What do we mean by asking whether economics is a science?, in: Alfred S. Eichner, eds., Why economics is not yet a science (MacMillan Press, London) 15–40.
Carnap, Rudolph, 1952, The continuum of inductive methods (University of Chicago Press, Chicago, IL).
De Marchi, Neil, ed., 1988, The Popperian legacy in economics (Cambridge University Press, Cambridge).
Earman, John, 1983, Testing scientific theories, Minnesota studies in the philosophy of science (University of Minnesota Press, Minneapolis, MN).
Earman, John, 1992, Bayes or bust? (MIT Press, Cambridge, MA).
Engle, Robert F., 1984, Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam) 775–826.
Fisher, R.A., 1955, Statistical methods and scientific induction, Journal of the Royal Statistical Society B 17, 69–78.
Fisher, R.A., 1973a, Statistical methods for research workers, 14th ed. (Hafner Publishing Company, New York, NY).
Fisher, R.A., 1973b, Statistical methods and scientific inference, 3rd ed. (Hafner Press, New York, NY).
Fisher Box, Joan, 1978, R.A. Fisher, The life of a scientist (Wiley, New York, NY).
Friedman, Milton, 1940, Review of Tinbergen (1939), American Economic Review 30, 657–661.
Friedman, Milton, 1988, Money and the stock market, Journal of Political Economy 96, 221–239.
Friedman, Milton and Anna J. Schwartz, 1991, Alternative approaches to analyzing economic data, American Economic Review 81, 39–49.
Giere, Ronald N., 1983, Testing theoretical hypotheses, in: Earman (1983, 269–298).
Giere, Ronald N., 1988, Explaining science, a cognitive approach (University of Chicago Press, Chicago, IL).
Godfrey, L.G., 1988, Misspecification tests in econometrics (Cambridge University Press, Cambridge).
Goldberger, Arthur S., 1991, A course in econometrics (Harvard University Press, Cambridge, MA).
Granger, Clive, Maxwell L. King, and Halbert White, 1995, Testing economic theories and the use of model selection criteria, Journal of Econometrics, this issue.
Haavelmo, Trygve, 1944, The probability approach in econometrics, Econometrica 12, Suppl.
Hacking, Ian, 1975, The emergence of probability (Cambridge University Press, Cambridge).
Hacking, Ian, 1983, Representing and intervening: Introductory topics in the philosophy of natural science (Cambridge University Press, Cambridge).
Hahn, Frank, 1992, Answer to Backhouse, RES Newsletter no. 78, July.
Hall, P. and B. Selinger, 1986, Statistical significance: Balancing evidence against doubt, Australian Journal of Statistics 28, 354–370.
Hardle, Wolfgang and Alan Kirman, 1995, Nonclassical demand: A model-free examination of price quantity relations in the Marseille fish market, Journal of Econometrics, this issue.
Hendry, David F., 1980, Econometrics: Alchemy or science?, Economica 47, 387–406.
Hendry, David F., 1992, Assessing empirical evidence in macroeconometrics with an application to consumers' expenditure in France, in: Allessandro Vercelli and Nicola Dimitri, eds., Macroeconomics: A survey of research strategies (Oxford University Press, Oxford) 363–392.

Hendry, David F. and Neil R. Ericsson, 1991, An econometric analysis of U.K. money demand in *Monetary Trends in the United States and the United Kingdom* by Milton Friedman and Anna J. Schwartz, American Economic Review 81, 8–38.
Howson, Colin and Peter Urbach, 1989, Scientific reasoning: The Bayesian approach (Open Court, La Salle).
Jeffreys, Harold, 1961, Theory of probability, 3rd ed. (Clarendon Press, Oxford).
Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, 1988, Introduction to the theory and practice of econometrics, 2nd ed. (Wiley, New York, NY).
Kendall, M.G., 1963, Ronald Aylmer Fisher, 1890–1962, Biometrika 50, 1–15.
Keuzenkamp, Hugo A., 1994, Probability, econometrics and truth: A treatise on the foundations of econometric inference, Unpublished Ph.D. thesis (Department of Economics, Tilburg University, Tilburg).
Keuzenkamp, Hugo A. and Anton P. Barten, 1995, Rejection without falsification, on the history of testing the homogeneity condition in the theory of consumer demand, Journal of Econometrics, this issue.
Keuzenkamp, Hugo A. and Michael McAleer, 1995, Simplicity, scientific inference, and econometric modelling, Economic Journal 105, Jan.
Keynes, John Maynard, 1921, A treatise on probability: The collected writings of John Maynard Keynes, VIII (St. Martin's Press, New York, NY).
Koopmans, Tjalling, 1937, Linear regression analysis of economic time series (De Erven F. Bohn, Haarlem).
Koopmans, Tjalling, 1947, Measurement without theory, Review of Economic Statistics 29, 161–172.
Lakatos, Imre, 1978, Philosophical volume papers 1: The methodology of scientific research programmes (Cambridge University Press, Cambridge).
Leamer, Edward E., 1978, Specification searches (Wiley, New York, NY).
Lehmann, E.L., 1986, Testing statistical hypotheses, 2nd ed. (Wiley, New York, NY).
Lehmann, E.L., 1993, The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two?, Journal of the American Statistical Association 88, 1242–1249.
McCloskey, Donald, 1985, The rhetoric of economics (University of Wisconsin Press, Madison, WI).
Mirowski, Philip, 1995, Three ways to think about testing in econometrics, Journal of Econometrics, this issue.
Neyman, Jerzy and Egon S. Pearson, 1928, On the use and interpretation of certain test criteria for purposes of statistical inference, I & II, Biometrika 20 A, 175–200, 263–294.
Neyman, Jerzy and Egon S. Pearson, 1933, On the problem of the most efficient test of statistical hypotheses, Philosophical Transactions of the Royal Society A 231, 289–337.
Pearson, Karl, 1900, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine 50, 157–172.
Popper, Karl R., 1968, The logic of scientific discovery, 2nd ed. (Harper & Row, New York, NY).
Reid, Constance, 1982, Neyman—From life (Springer-Verlag, Berlin).
Savage, Leonard J., 1961, The foundations of statistics reconsidered, in: Jerzy Neyman, ed., Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol. I (University of California Press, Berkeley, CA).
Savage, Leonard J., 1972, The foundations of statistics, 2nd rev. ed. (Dover, New York, NY).
Spanos, Aris, 1986, Statistical foundations of econometric modelling (University of Cambridge Press, Cambridge).
Spanos, Aris, 1995, On theory testing in econometrics: The case of the efficient market hypothesis, Journal of Econometrics, this issue.
Stigler, George J., 1965, Essays in the history of economics (University of Chicago Press, Chicago, IL).

Stigum, Bernt P., 1990, Toward a formal science of econometrics (MIT Press, Cambridge, MA).
Summers, Lawrence H., 1991, The scientific illusion in empirical macroeconomics, Scandinavian Journal of Economics 93, 129–148.
Theil, Henri, 1971, Principles of econometrics (Wiley, New York, NY).
Tinbergen, Jan, 1939, Statistical testing of business cycle theories, Two vols. (League of Nations, Economic Intelligence Service, Geneva).
Vuong, Quang H., 1989, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57, 307–333.
Wonnacott, Ronald J. and Thomas H. Wonnacott, 1985, Introductory statistics, 4th ed. (Wiley, New York, NY).
Zellner, Arnold, 1979, Posterior odds ratios for regression hypotheses: General considerations and some results, Reprinted in: Basic issues in econometrics (University of Chicago Press, Chicago, IL) 275–305.
Zellner, Arnold, 1988, Bayesian analysis in econometrics, Journal of Econometrics 37, 27–50.

INSTRUCTIONS TO AUTHORS

(1) Papers must be in English.

(2) Papers for publication, accompanied by a submission fee of $50.00 for authors who currently do not subscribe to this journal (subscribers are exempt), should be sent in quintuplicate to:
Professor Cheng Hsiao, Department of Economics, University of Southern California, Los Angeles, CA 90089, USA.
Submission of a paper will be held to imply that it contains original unpublished work and is not being submitted for publication elsewhere. The Editor does not accept responsibility for damage or loss of papers submitted. Upon acceptance of an article, author(s) will be asked to transfer copyright of the article to the publisher. This transfer will ensure the widest possible dissemination of information.

(3) Submission of accepted papers as *electronic manuscripts*, i.e., on disk with accompanying manuscript, is encouraged. Electronic manuscripts have the advantage that there is no need for rekeying of text, thereby avoiding the possibility of introducing errors and resulting in reliable and fast delivery of proofs. The preferred storage medium is a 5.25 or 3.5 inch disk in MS-DOS format, although other systems are welcome, e.g., Macintosh (in this case, save your file in the usual manner; do not use the option "save in MS-DOS format"). Do not submit your original paper as electronic manuscript but hold on to the disk until asked for this by the Editor (in case your paper is accepted without revisions). Do submit the accepted version of your paper as electronic manuscript. Make absolutely sure that the file on the disk and the printout are identical. Please use a new and correctly formatted disk and label this with your name; also specify the software and hardware used as well as the title of the file to be processed. Do not convert the file to plain ASCII. Ensure that the letter "l" and digit "1", and also the letter "O" and digit "0" are used properly, and format your article (tabs, indents, etc.) consistently. Characters not available on your word processor (Greek letters, mathematical symbols, etc.) should not be left open but indicated by a unique code (e.g., gralpha, ⟨alpha⟩, @, etc., for the Greek letter α). Such codes should be used consistently throughout the entire text; a list of codes used should accompany the electronic manuscript. Do not allow your word processor to introduce word breaks and do not use a justified layout. Please adhere strictly to the general instructions below on style, arrangement, and, in particular, the reference style of the journal.

(4) Manuscripts should be double spaced, with wide margins, and printed on one side of the paper only. All pages should be numbered consecutively. Titles and subtitles should be short. References, tables, and legends for figures should be printed on separate pages.

(5) The first page of the manuscript should contain the following information: (i) the title; (ii) the name(s) and institutional affiliation(s) of the author(s); (iii) an abstract of not more than 100 words. A footnote on the same sheet should give the name, address, telephone number, fax number, and E-mail address of the corresponding author.

(6) The first page of the manuscript should also contain at least one classification code according to the Classification System for Journal Articles as used by the *Journal of Economic Literature*; in addition, up to five key words should be supplied.

(7) Acknowledgements and information on grants received can be given in a first footnote, which should not be included in the consecutive numbering of footnotes.

(8) Footnotes should be kept to a minimum and numbered consecutively throughout the text with superscript Arabic numerals. They should be double spaced and not include displayed formulae or tables.

(9) Displayed formulae should be numbered consecutively throughout the manuscript as (1), (2), etc. against the right-hand margin of the page. In cases where the derivation of formulae has been abbreviated, it is of great help to the referees if the full derivation can be presented on a separate sheet (not to be published).

(10) References to publications should be as follows: "Smith (1992) reported that ..." or "This problem has been studied previously (e.g., Smith et al., 1969)". The author should make sure that there is a strict one-to-one correspondence between the names and years in the text and those on the list. The list of references should appear at the end of the main text (after any appendices, but before tables and legends for figures). It should be double spaced and listed in alphabetical order by author's name. References should appear as follows:
For monographs
Hawawini, G. and I. Swary, 1990, Mergers and acquisitions in the U.S. banking industry: Evidence from the capital markets (North-Holland, Amsterdam).
For contributions to collective works
Brunner, K. and A.H. Meltzer, 1990, Money supply, in: B.M. Friedman and F.H. Hahn, eds., Handbook of monetary economics, Vol. 1 (North-Holland, Amsterdam) 357–396.
For periodicals
Griffiths, W. and G. Judge, 1992, Testing and estimating location vectors when the error covariance matrix is unknown, Journal of Econometrics 54, 121–138.
Note that journal titles should not be abbreviated.

(11) Illustrations will be reproduced photographically from originals supplied by the author; they will not be redrawn by the publisher. Please provide all illustrations in *quadruplicate* (*one* high-contrast original and *three* photocopies). Care should be taken that lettering and symbols are of a comparable size. The illustrations should not be inserted in the text, and should be marked on the back with figure number, title of paper, and author's name. All graphs and diagrams should be referred to as *figures*, and should be numbered consecutively in the text in Arabic numerals. Illustrations for papers submitted as electronic manuscripts should be in traditional form.
Illustrations can be printed in color when they are judged by the Editor to be essential to the presentation. The publisher and the author will each bear part of the extra costs involved. Further information concerning color illustrations and the costs to the author can be obtained from the publisher.

(12) Tables should be numbered consecutively in the text in Arabic numerals and printed on separate sheets.

Any manuscript which does not conform to the above instructions may be returned for the necessary revision before publication.

Page proofs will be sent to the corresponding author. Proofs should be corrected carefully; the responsibility for detecting errors lies with the author. Corrections should be restricted to instances in which the proof is at variance with the manuscript. Extensive alterations will be charged. Twenty-five reprints of each paper are supplied free of charge to the corresponding author; additional reprints are available at cost if they are ordered when the proof is returned.